Brandilyn Hall, Elyssa Irizarry, Megan Adams, Yvonne Martinez
Data Science Boot Camp
Project 4 - Group 1

Housing Predictions

The Real Estate Market has been “running hot” for quite some time with no end in sight. Houses sell within a day for often times more than it was listed which, in turn, makes it the true definition of a sellers’ market. While this type of market looks enticing to potential sellers, selling a house is a big decision. Our project was built with potential house sellers in mind. We created a Housing prediction page that uses Machine Learning to predict house price from house information the potential seller enters. In this report, we will introduce you to our data and walk you through our analysis, tableau design, and our machine learning process. We will also discuss our web application development, design, deployment, and navigation.

Our dataset came from the NTREIS[1] (North Texas Real Estate Information Service) website which is used by realtors and contains housing data for the DFW area. We were able to pull the most recent 5,000 rows of information on houses sold within the last 6 months in Dallas, Collin, Ellis, Hunt, Kaufman, Rockwall, and Tarrant Counties. We were also able to pull many attributes of each house sold including, number of bedrooms and bathrooms, year built, garage included, fireplace included, close date, and many more. This data along with the multitude of features should aid in helping us predict closing prices of other homes in the area. We also pulled Census[2] data on the DFW counties mentioned above to use for some of our Tableau visualizations.

Our inspiration for this topic came from our interest in the climate of the market currently. We began looking at other housing prediction projects that others had started on Kaggle for additional information (Austin Housing[3], House Price Prediction[4]) as well as TAMU’s Housing Dashboard[5] and Tableau[6] for visualization ideas. We took ideas from all of these sources to create our own housing prediction dashboard as well as Tableau visualizations.

From our review of our dataset, we propose the following research questions:

- Which area of DFW has the fastest sale rate?
- Which area of DFW sells houses for the most money?
- Which features contribute to making a house sell faster and for more money?

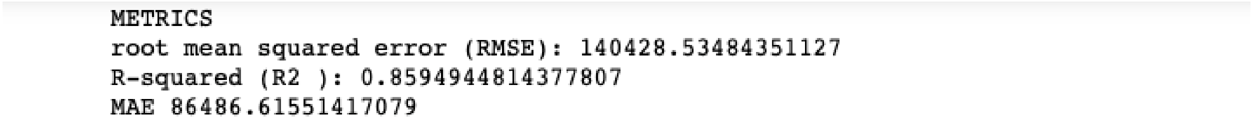
From our prior knowledge of the real estate market as well as our review of our inspiration resources, we propose the following hypothesis: *Location, square footage, number of bed/bathrooms, and additions (e.g, garage, pools, sheds) affects the price and time on the market. Essentially, the more of these features, the higher the price.*

Machine Learning

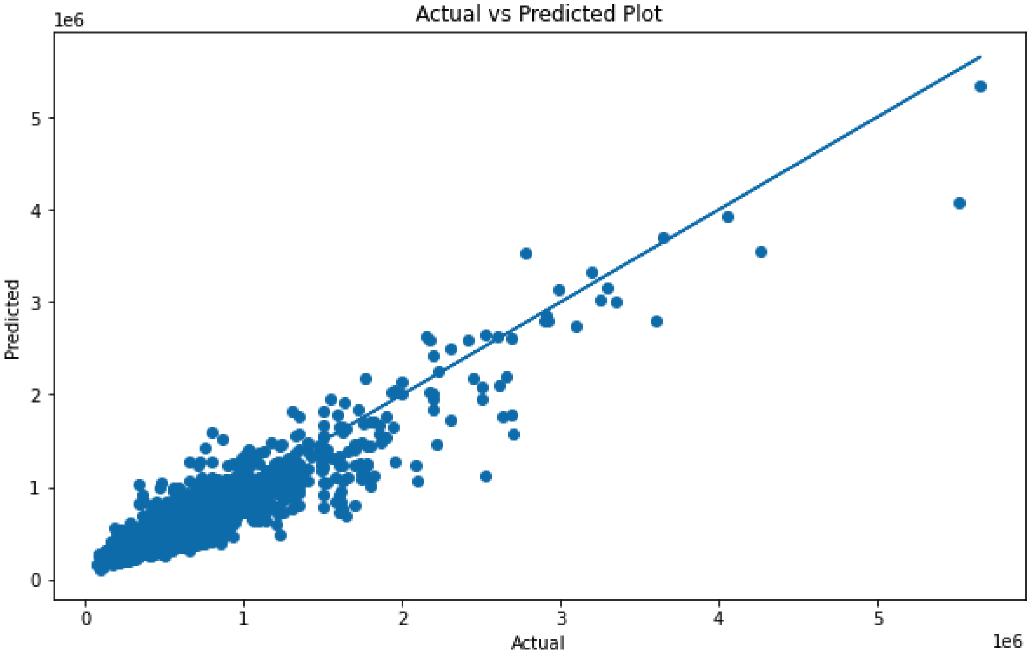
[1] - https://ntreis.net/accessing-ntreis-data/accessing-ntreis-data-agents/
[2] - Texas - Census Bureau Tables
[3] - https://www.kaggle.com/code/threnjen/austin-housing-eda-nlp-models-visualizations
[4] - https://www.kaggle.com/code/kenywhite/house-price-prediction
[5] - https://www.recenter.tamu.edu/data/housing-dashboard/
[6] - https://public.tableau.com/app/profile/us.census.bureau/viz/2020CensusPopulationandHousingMap/AllStates

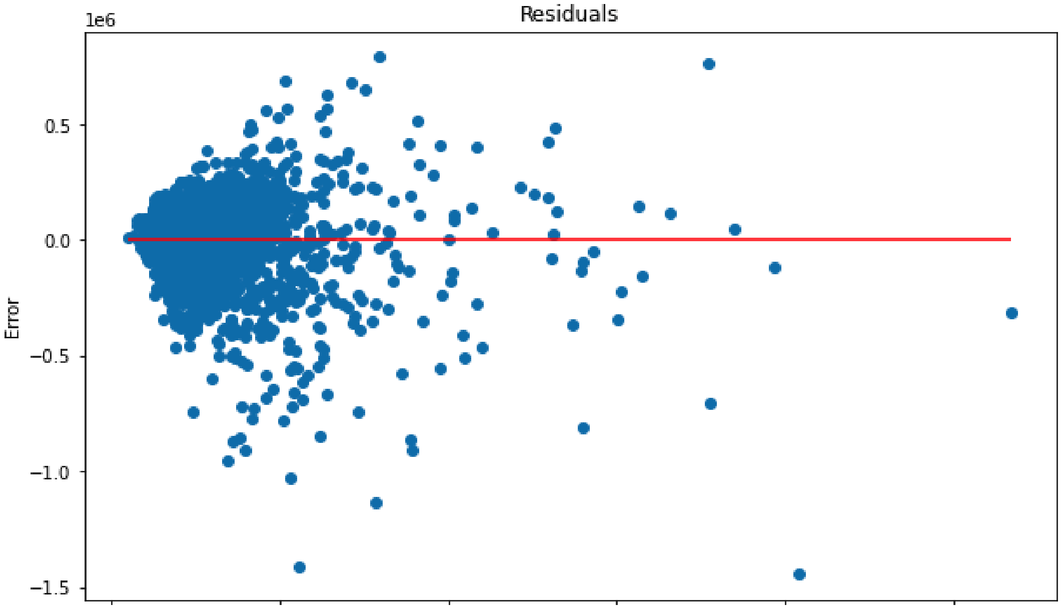
Using the NTREIS data of the Dallas/Fort Worth area from the past months we were able to create a prediction of the closing price of future home sales. Supervised machine learning was performed on the data. The five thousand line data was split into training and test sets at seventy five and twenty five percent, respectfully. The data was randomly assigned to the two sets. Columns that duplicated information such as HOA Fees, bathroom amounts, and bedroom amounts. Address and subdivision names were also dropped as features. Null values were removed. Numeric columns were converted to integer. Closing date was changed to datetime and removed from the data set to be brought back into the data as needed in the future. County and HOA type was one hot encoded. City, heating, and utilities were label encoded. It was determined that the numeric data would not be scaled based upon the relation of the dataset. Features correlation was ranked to determine most relevant features for the closing cost prediction. The training data was then split to run training and testing through regression models.

Multiple regression models were run including linear regression, ridge, lasso, elastic net, decision tree, random forest, extra trees, ada boost, gradient boost, and xbg regressor. Features with the strongest correlation were run through the models. After model attempts, the following features were chosen to run the best model: county, city, hoa type, square feet, bedroom total, and bathroom total. The best model was determined to be gradient boost due to the accuracy and similarities between the training and testing numbers. The training data showed an r-squared of 0.867 and the testing data;s r-squared was 0.74. The model was then run on the entire training set to create a final prediction model.



```
METRICS
root mean squared error (RMSE): 140428.53484351127
R-squared (R2 ): 0.8594944814377807
MAE 86486.61551417079
```

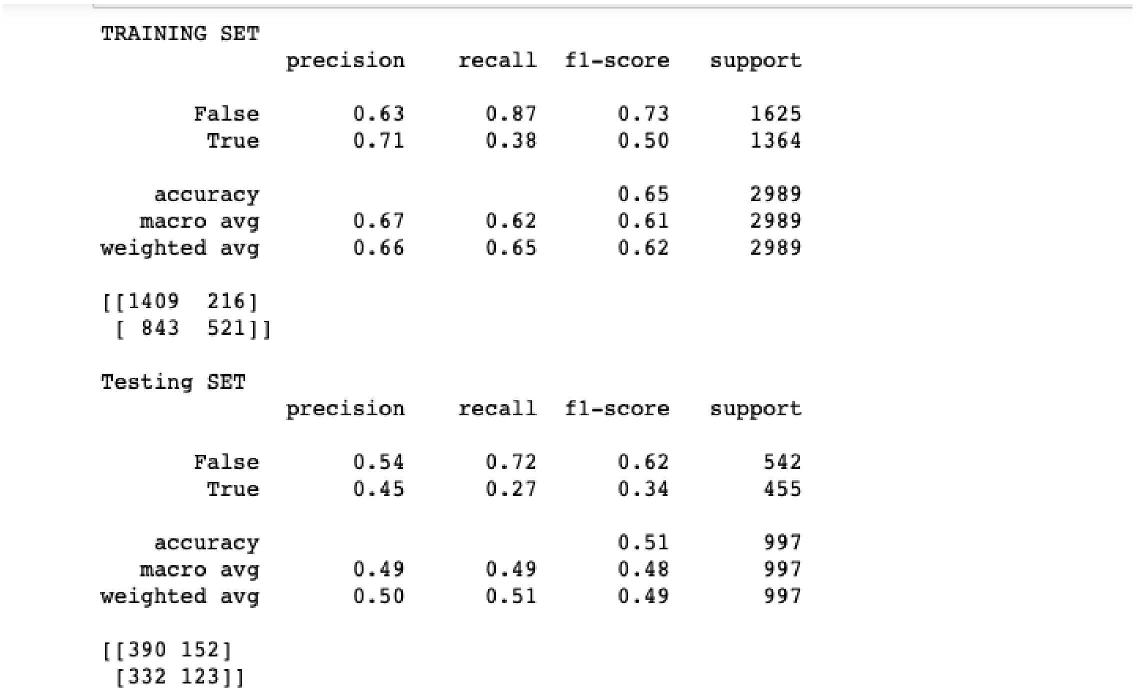




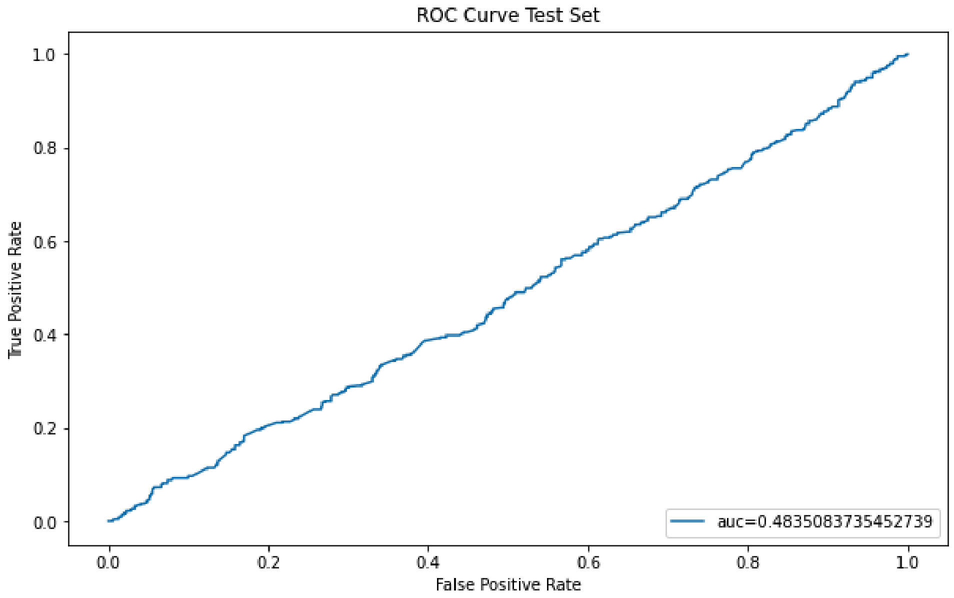
The remaining testing data from the original housing data was loaded, cleaned, and encoded identically to the training data. The model was loaded and ran over the testing data. The accuracy of the model averaged 20.12% in relation to actual closing prices. However, the accuracy of the prediction ranged from 0.01% to 265.5% accuracy. The data was predicted at

32.5% accuracy most often. This is a wide range of predictability, showing that our model would benefit from more data or different data to be used as features in order to more accurately predict home closing prices.

An attempt at a second prediction model was attempted to see if the days a home would be on the market could be determined. The training data was cleaned and encoded as in the first prediction model. The same series of regressions models were run on the data with the same features in the first model. The data showed no predictable correlation. The features were decreased to see if accuracy could be gained. No increase in predictable correlation was shown. Upon further analysis of the days on market target, it was determined that the majority of homes were sold in less than a week. A boolean category was created to denote if a home was sold in less or more than five days. This target was then run through a series of classification models to determine if a predictable correlation could be found. The model was determined as not applicable and was not used in our final prediction.

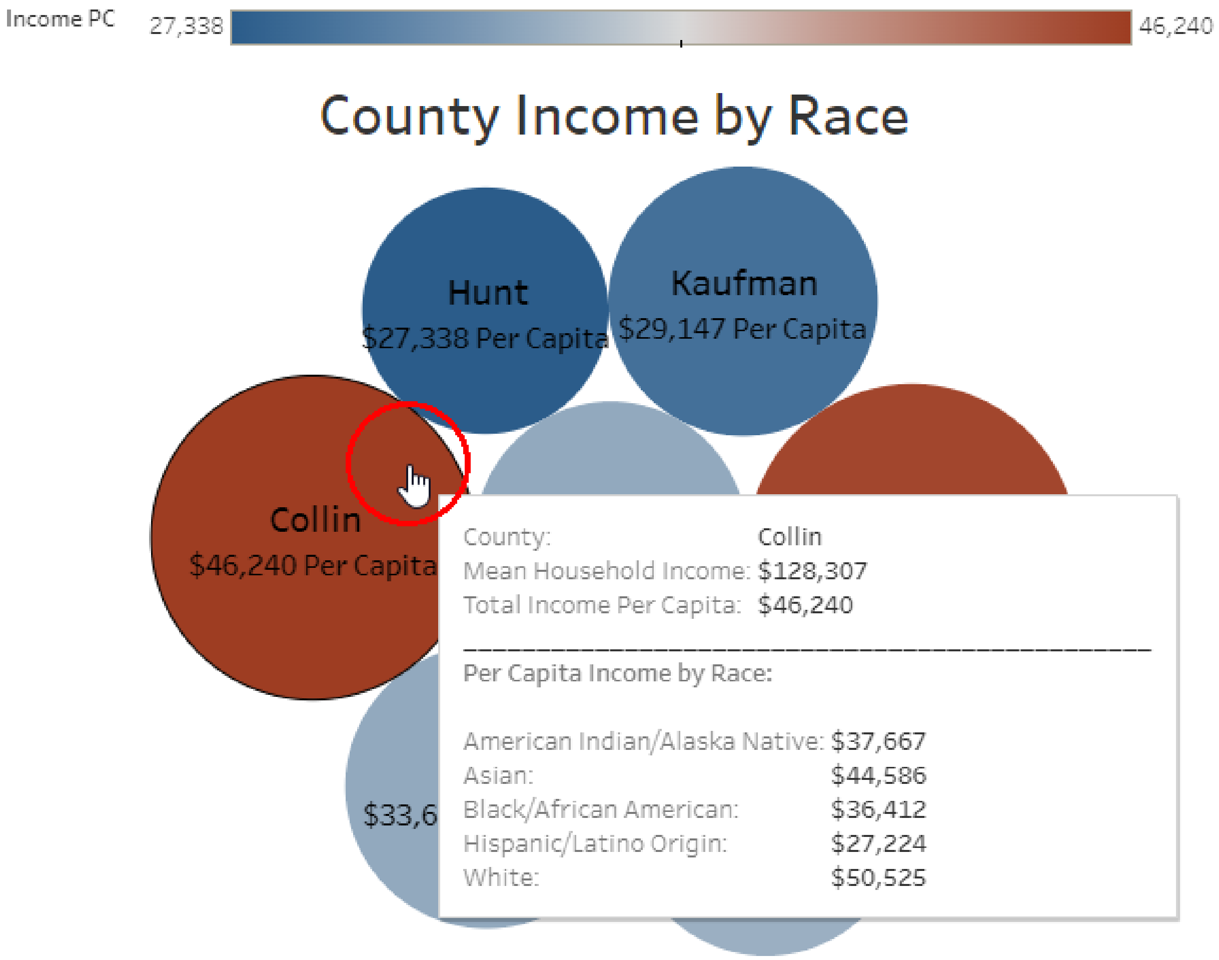


```
TRAINING SET
              precision    recall  f1-score   support

       False       0.63      0.87      0.73      1625
        True       0.71      0.38      0.50      1364

    accuracy                           0.65      2989
   macro avg       0.67      0.62      0.61      2989
weighted avg       0.66      0.65      0.62      2989

[[1409  216]
 [ 843  521]]

Testing SET
              precision    recall  f1-score   support

       False       0.54      0.72      0.62       542
        True       0.45      0.27      0.34       455

    accuracy                           0.51       997
   macro avg       0.49      0.49      0.48       997
weighted avg       0.50      0.51      0.49       997

[[390 152]
 [332 123]]
```



Tableau

Using Tableau Public, we created three separate Dashboards. The first one focused on 2020 census data. For each county, we pulled several CSVs covering a variety of topics: race, education, age, income, and population. From there, a read-only census dashboard was formed.

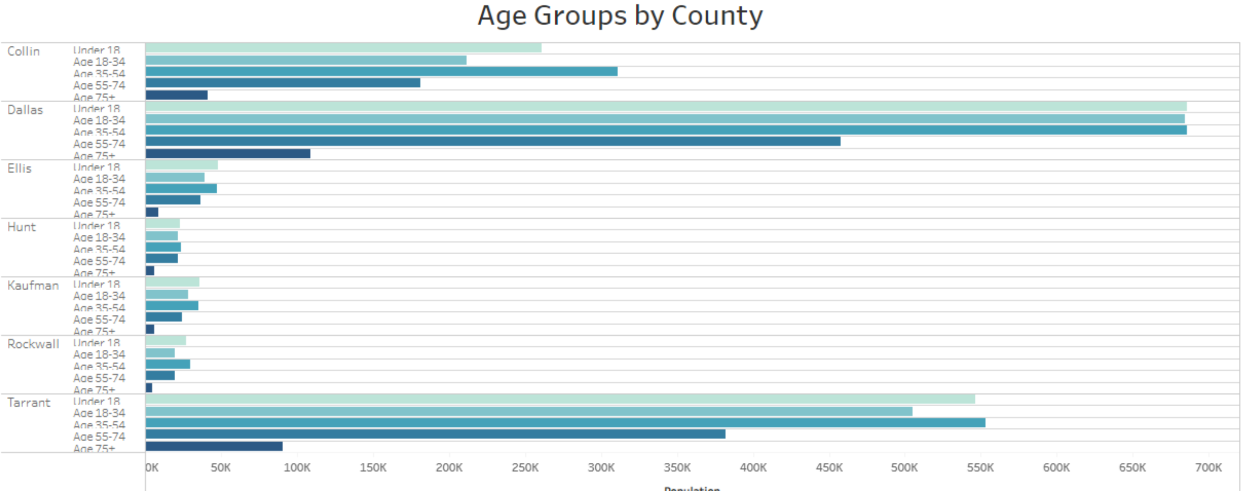


County: Collin
Mean Household Income: $128,307
Total Income Per Capita: $46,240

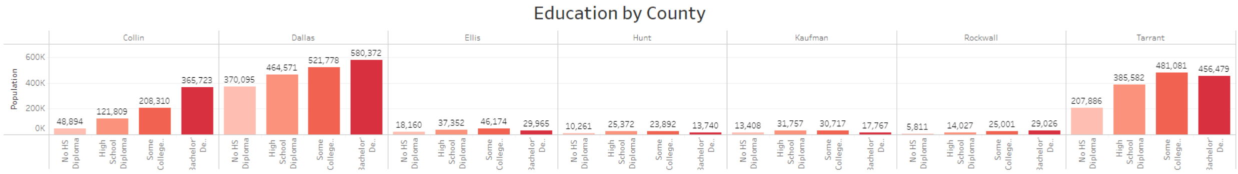
Per Capita Income by Race:

American Indian/Alaska Native: $37,667
Asian: $44,586
Black/African American: $36,412
Hispanic/Latino Origin: $27,224
White: $50,525

One bubble graph compared the races in each county by their per capita income, as well as their mean household income. It's grouped by county and color-coded by lowest to highest per capita income. Collin and Rockwall counties had the highest overall per capita income by far, more than $10,000 higher than next in line, Tarrant. They also had the highest mean household income at $128,307 and $133,000 respectively, the only ones in the six-digits. This higher bracket is reflected across all races except Asians and American Indians/Alaska Natives, where those groups did better in Dallas and Kaufman.

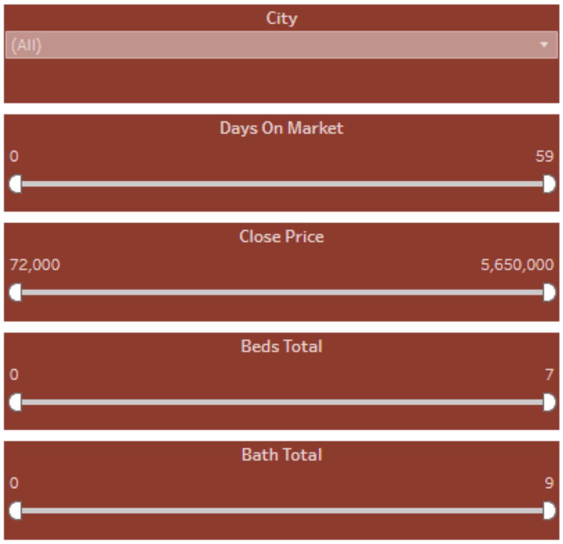


Another horizontal bar graph compared age groups for each county. This one required the most amount of calculated fields, because the raw census data was specific to a sex and much smaller age group. For example, the “Age 18-34” field sums up “18 and 19,” “20,” “21”, “22 to 24,” “25 to 29,” and “30 to 34,” which were also originally split out by sex. So twelve fields went into this one calculated age group. Analyzing the graph shows that the most young people (under 35) live in Dallas County, with Tarrant County a close second. Interestingly, even with their large population, Tarrant County has more people under 18 and ages 35-54 than ages 18-34. Either, it looks like once a young person reaches adulthood, many leave the county - or the ones living there are having many children to outpace the next age group.



The final census graph compares education across each county. The census data was split out by age groups and types of degrees, such as “25+ Associate’s degree.” We created calculated fields breaking the categories down into more easily read groups: “No high school diploma,” “high school diploma,” “some college or Associate’s,” and “Bachelor’s degree or higher.” This is a vertical bar graph summed up by population. It shows that Dallas County has both the most people with a degree as without, due to their large population. Even though Rockwall has a higher per capita income, Ellis has more total people with a Bachelor’s degree.

Two dashboards were created out of the sold home dataset. Both the map and the graphs have the ability to filter the days on market, closing price, bedroom total, and bathroom total. We included the city as a filter for the Sold Homes dashboard as well.



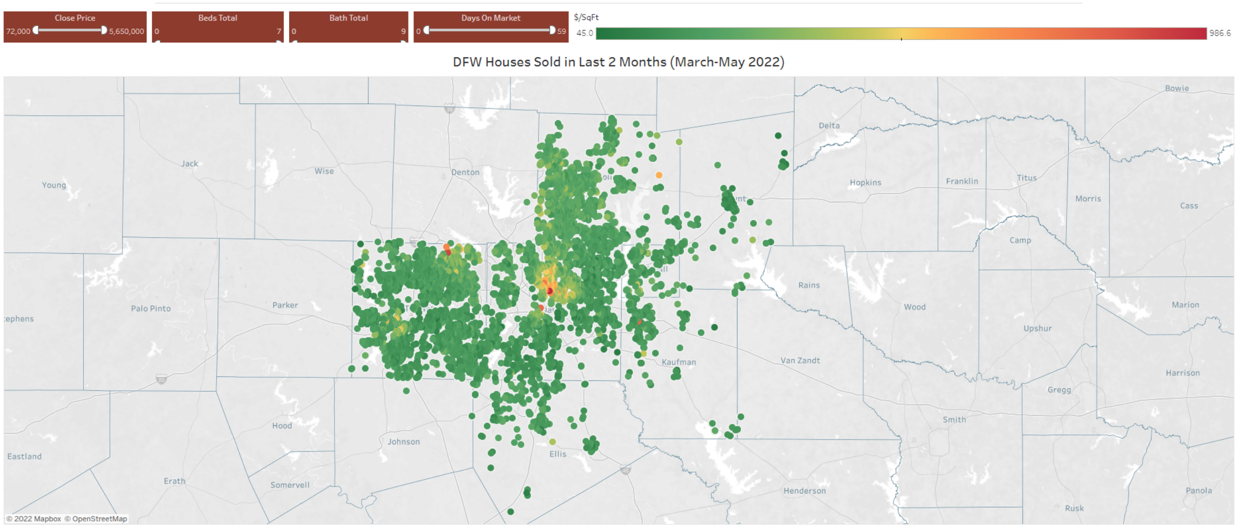
City
(All)

Days On Market
0 59

Close Price
72,000 5,650,000

Beds Total
0 7

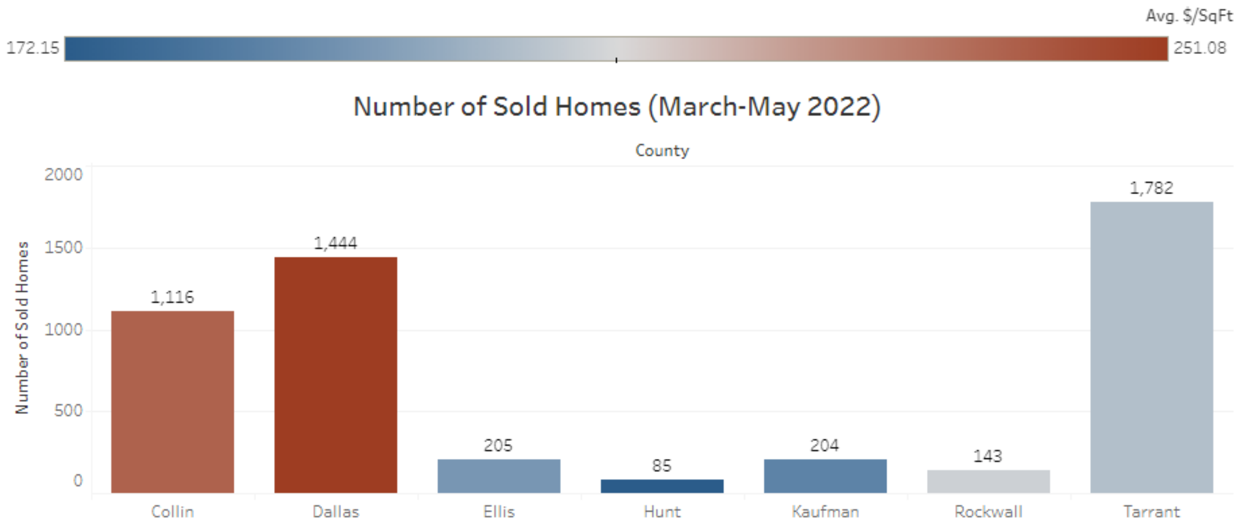
Bath Total
0 9



Close Price 72,000 5,650,000
Beds Total 0 7
Bath Total 0 9
Days On Market 0 59
$/SqFt 45.0 986.6

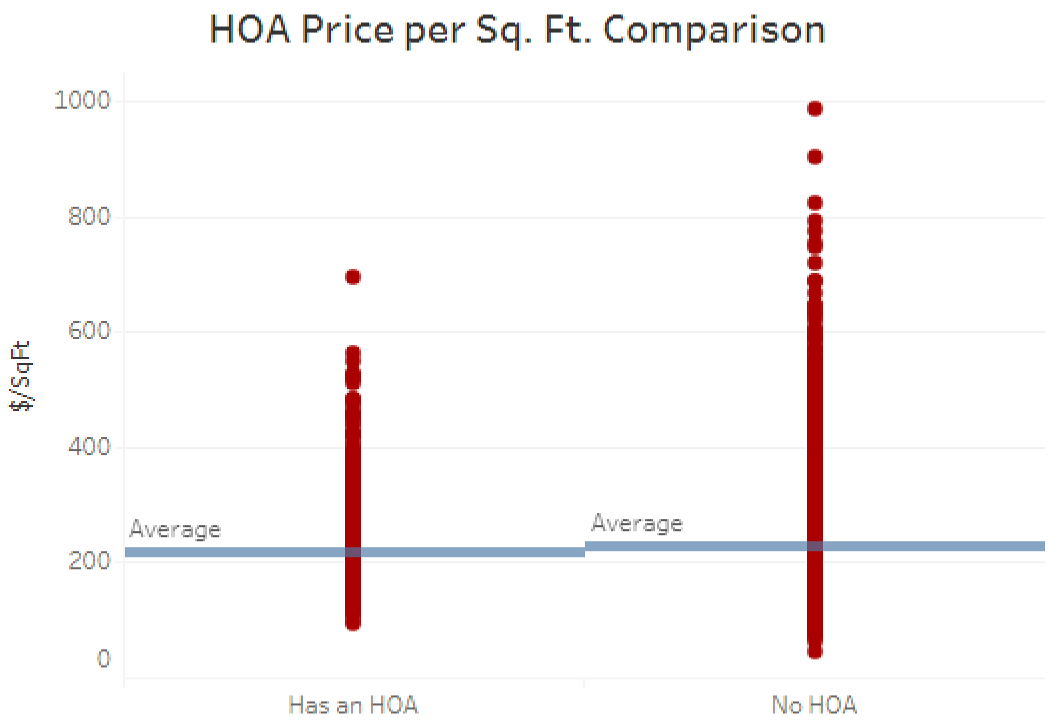
DFW Houses Sold in Last 2 Months (March-May 2022)

The first dataset dashboard is a map of almost all 5,000 rows of data, with each location color-coded by the closing price per square foot. Exclusions were a few outlier homes that threw off the coloring for the entire map. The tooltip included the address, closing data, closing price, total square footage, price per square foot, and days on market. It showed that the most expensive homes per square foot are squarely in the middle of Dallas, around Highland Park, as well as out by Southlake.

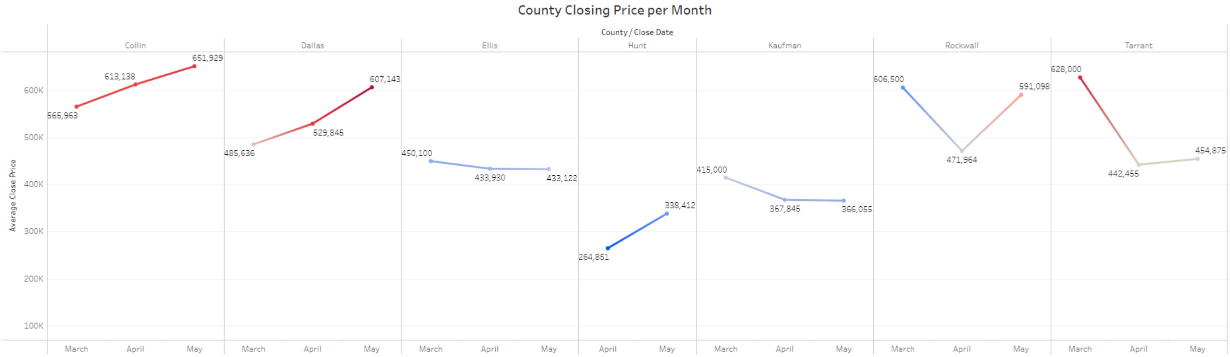


County Closing Price per Month

The second dashboard included three graphics. The first is a vertical bar graph showing the number of sold homes grouped by county, color-coded by price per square foot. Dallas and Collin had the highest price per square foot, while Tarrant had the lowest. That may have been a motivating factor for buyers, because Tarrant had most of the homes sold in the dataset.



The second is a box-and-whisker plot grouping each sold home by whether or not they have an HOA, with the y-axis being price per square foot. This shows the average price per square foot is actually higher in homes without an HOA. This looks to be because the majority of the higher-priced homes without an HOA are in Highland Park and Southlake, one of the most expensive areas of DFW per the map, dragging the average upward.



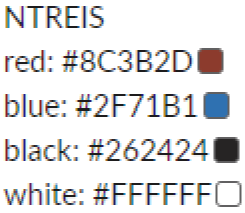
The last graph shows the average county's closing price separated by month, color-coded by price per square foot. Between March and May 2022, closing prices overall went up in Collin, Dallas, and Hunt counties, and down in Ellis, Kaufman, Rockwall, and Tarrant counties. However, Kaufman and Tarrant counties were the only ones to have the price per square foot drop. Everywhere else, that increased, sometimes dramatically like in Dallas ($207 to $256/sq. ft).

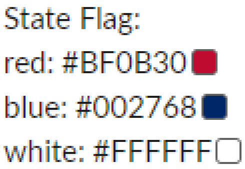
Web App Section

Our goal was to create a clean user friendly website, where we can showcase all the skills we have developed thus far in the bootcamp. We have a total of 9 pages which include : Home, Tableau dashboard on housing data,Tableau dashboard on DFW area, Tableau dashboard on census data, Prediction Model, Writeup, Abous Us, Data Source, and Work Cited.

We used a design template from BootSwatch for the HTML, as well as added more styling through our CSS stylesheet. This allowed for more personalization to our website.

Our color design is intentional and took inspiration from our state flag and NTREIS logo colors. Overall our design was meant to stay clean and organized to not overwhelm a user and really focus on the data.


NTREIS
red: #8C3B2D🟫
blue: #2F71B1🟦
black: #262424⬛
white: #FFFFFF⬜


State Flag:
red: #BF0B30🟥
blue: #002768🟦
white: #FFFFFF⬜

Conclusion/Call to Action

Based on our analysis, we came to the following conclusions:

- The more “features” a house has (e.g., rooms, garage, etc.,) the more the house will sell for.
- Surprisingly, houses that had an HOA were not more expensive the non-HOA houses
- The location of the house affects the house price. Houses in more affluent areas with higher education recipients typically cost more.

Limitations and Future Work

We discovered a few limitations as we were working on our project which did not allow us to achieve all of the objectives we had hoped. One of the limitations was the NTREIS dataset captured only DFW information which narrowed down the scope of our analysis and predictions. Also, the NTREIS dataset only allowed us to pull the most current 5,000 rows of data of the last 6 months. Due to the limit on rows, we were not able to view many sold houses in January or February 2022. Finally, the biggest limitation we encountered was our “days on market” target was not able to be predicted. When we ran models to predict days on market, we received negative r-squared values which indicates a very poor prediction. No matter what we tried, we were unable to get a solid prediction, so we eventually had to decide to drop that prediction.

In the future we would like to expand our scope of analysis by obtaining data from other areas of Texas and the United States. This will not only give us more rows of data but also allow us to predict selling prices for more potential sellers. With more data and possibly more features, we hopefully also be able to predict days on market in the future, which would help sellers make more informed decisions on when to list their house. This will require us using additional real estate databases to gather more inclusive data, as well as data on rental properties in the areas reviewed.

Conclusion

After the selection of our features and models, we were able to successfully predict the closing price of houses in the DFW area. The machine learning along with our Tableau visualizations and designed web app not only allows the user to learn the predicted price of the house they want to sell, but also tells the story of the housing data in the DFW area. The user will be able to input their own data and interact with data to learn more about the attributes that contribute to the value of their home.